Faculté des Sciences
Aix*Marseille Université

Année universitaire 2022/2023

Site : ☐ Luminy ☒ St-Charles ☐ St-Jérôme ☐ Cht-Gombert ☐ Aix-Montperrin ☐ Aubagne-SATIS
Sujet de : ☒ 1er semestre ☐ 2ème semestre ☐ Session 2 Durée de l'épreuve : 2h
Examen de : M1 Nom du diplôme : Master MAS
Code du module : SMSA05AC Libellé du module : Statistique exploratoire.
Calculatrices autorisées : OUI Documents autorisés : OUI, supports de cours

---

Une étude (les données sont présentées en Annexe 1) a été menée sur 41 villes des États-Unis sur lesquelles 6 variables ont été observées :

- **Temp.** : température moyenne annuelle (en degrés Celsius) ;
- **Indus.** : nombre d'industries employant 20 salariés et plus ;
- **Pop.** : taille de la population (en milliers) ;
- **Vent** : vitesse moyenne du vent sur une année (en km/h) ;
- **Precip.** : précipitation moyenne annuelle (en cm) ;
- **Jour** : nombre moyen de jours de pluie par an.

L'objectif de cette étude consiste à réaliser une ACP sur ces 6 variables afin d'en extraire de l'information.

A - **Phase exploratoire**

Avant de se lancer directement dans une ACP, il est souvent judicieux de mener préalablement un travail exploratoire des données. L'intérêt d'un tel travail est d'accumuler un maximum d'information tant d'un point de vue univarié que bivarié, et ce afin de pouvoir proposer a posteriori d'éventuelles interprétations aux résultats de l'ACP.

1) Compléter le tableau suivant :

	Min	Q_1	Med	Moyenne	Q_3	Max	Écart-type	CV
Temp.	6.40	10.30	12.60	13.20	15.20	24.20	3.97	0.30
Indus.	35.00	181.00	347.00	463.10	462.00	···	556.56	···
Pop.	71.00	299.00	515.00	608.60	717.00	···	572.01	0.94
Vent	9.70	14.00	15.00	15.20	17.10	20.40	2.27	···
Precip.	17.91	78.64	98.40	93.39	109.50	151.89	29.53	0.32
Jour	···	103.00	115.00	113.90	128.00	166.00	26.18	0.23

TABLE 1 – Statistiques usuelles

Note : on rappelle que la statistique CV désigne le *coefficient de variation* qui se calcule comme le rapport de l'écart-type par la moyenne.

2) À partir de la lecture de cette table, indiquer les variables qui présentent une forte dispertion. Justifier votre réponse, puis donner une intreprétation contextuelle.
3) Le graphique Fig. 1 fournit les boxplots des différentes variables. Que représentent les points symbolisés par des ronds ? Ces graphiques confirment-ils les commentaires efféctués aux questions précédentes ?
4) Le graphique Fig. 2 fournit les différents nuages de dispersion associés aux variables. Peut-on détecter à partir de ce graphique des corrélations linéaires ? Si oui, lesquelle(s) ? Justifier votre réponse.
5) Confirmer votre analyse précédente avec la matrice des corrélations donnée par la Table 2 (ci-dessous) :
6) Que signifie, dans le cadre de l'ACP, la dénomination **effet de taille** ? On sait qu'un effet de taille apparaîtra sur la première composante principale dès lors que la matrice

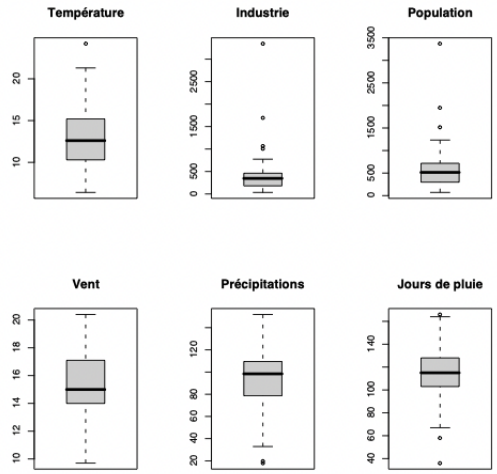

Fig. 1 : Boxplots

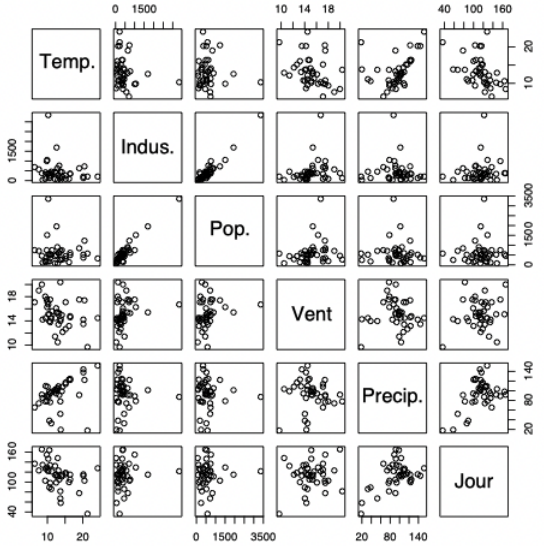

Fig. 2 : Nuages de dipersion

	Temp.	Indus.	Pop.	Vent	Precip.	Jour
Temp.	1.000	−0.19	−0.06	−0.35	0.38	−0.43
Indus.	−0.19	1.000	0.96	0.23	−0.03	0.13
Pop.	−0.06	0.96	1.000	0.21	−0.03	0.04
Vent	−0.35	0.23	0.21	1.000	−0.01	0.17
Precip.	0.38	−0.03	−0.03	−0.01	1.000	0.50
Jour	−0.43	0.13	0.04	0.17	0.50	1.000

TABLE 2 – Matrice des corrélations

des corrélations est “totalement positive”. Que pouvez-vous en déduire dans le cas présent ?

B - **Analyse en composantes principales**

La phase exploratoire étant achevée, il convient maintenant de réaliser une ACP sur le tableau de données sur les six variables.

1) Indiquer l’espace dans lequel seront représentées les variables. Justifier votre réponse.

2) Indiquer l'espace dans lequel seront représentés les observations sur lesquelles porte l'étude. Justifiez votre réponse.
3) Dans le cas présent, est-il préférable de mener une ACP normée ou une ACP standard (ou canonique) ? Justifier votre réponse ?
4) En faisant usage des statistiques usuelles (cf. Table 1), préciser les coordonnées du barycentre du nuage des observations, puis déterminer la matrice D_s. Quel est l'intérêt de calculer la matrice D_s dans le cas présent ?
5) D'après le cours, on sait que l'ACP normée consiste à diagonaliser la matrice des corrélations associée aux variables. Combien de valeurs propres doit-on trouver dans le cas présent ? Justifier votre réponse.
6) Dans le cadre de l'ACP normée, indiquer la relation reliant les valeurs propres au nombre de variables.
7) En faisant usage de la relation précédente, compléter le tableau suivant (où les inerties sont exprimées en pourcentage) :

λ_α	Inertie	Inertie cumulée
2.195	36.58	...
1.501	25.01	61.59
...	...	...
0.761	12.68	97.51
0.115	1.92	...
0.034	0.57	...

TABLE 3 – Valeurs propres et inertie

8) En se référant à la table des inerties (cf. Table 3), ainsi qu'au diagramme des valeur propres (cf. Fig. 3, ci-dessous), identifier le nombre de composantes principales que l'on sera amené à conserver. Justifier votre réponse.

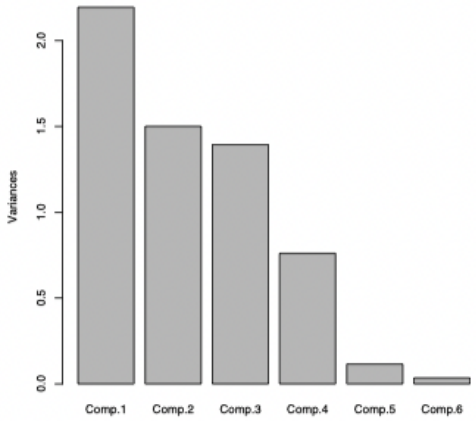

Fig. 3 : diagramme des valeurs propres

9) En vous aidant de l'Annexe 2, identifier les observations dont la contribution est supérieure à la contribution moyenne donnée en pourcentage par 100/41. Ces observations seront qualifiées de participant à la formation de l'axe. On réalisera un tableau du type suivant, où les symboles + et − représentent le signe de la coordonnée de l'observation sur l'axe correspondant.

Axe 1		Axe 2	
−	+	−	+
...	...	...	...
...	...	...	...

10) Selon la règle empirique "un axe factoriel est qualifié de robuste dès lors qu'au moins 10% des observations participent à sa formation", conclure sur la robustesse des deux premiers axes factoriels, en émettant éventuellement une critique.
11) Le graphique Fig. 4 représente les observations, autrement dit les villes, dont la qualité de représentation dans le plan est supérieure ou égale à 50%. Visualise-t-on des groupes de villes ? Si oui, lesquels ? Que signifie la proximité de deux villes (sous réserve d'une bonne qualité de représentation) ?
12) Rappeler la formule fondamentale du cours permettant d'obtenir directement la corrélation entre une variable x^j et une composante principale f^α.
13) Rappeler la formule permettant de calculer la qualité de représentation d'une variable sur un axe factoriel (noté Qual_α pour l'axe α) ainsi que sur un plan factoriel (noté $\text{Qual}_{\alpha,\beta}$ pour le plan α, β), puis compléter le tableau suivant :

	$a_{\cdot 1}$	$a_{\cdot 2}$	Qual_1	Qual_2	$\text{Qual}_{1,2}$
Temp.	−0.4902	0.1675	24.03	2.81	26.84
Indus.	0.9059	0.2121	...	4.50	86.56
Pop.	0.8558	0.2798	73.25	7.83	81.08
Vent	0.5217	−0.1668	...	...	...
Precip.	−0.0598	−0.7545	0.36	56.92	57.28
Jour	0.3544	−0.8674	12.56	75.24	87.80

14) En prenant comme seuil 50%, déterminer les variables corrélées avec chacun des deux premiers facteurs. Proposer, si possible, une interprétation contextuelle pour ces deux facteurs.
15) Est-il possible de visualiser directement à partir du nuage de points-variables la corrélation d'une variable avec un facteur ? Si oui, expliquer comment. Existe-t-il une condition à cette approche visuelle ?
16) Comment visualise-t-on la qualité de représentation d'une variable dans le premier plan factoriel ?
17) À l'aide du graphique Fig. 5, identifier les variables qui sont bien représentées dans le plan factoriel. Quelles sont les variables corrélées significativement avec le premier facteur ? Quelles sont les variables corrélées significativement avec le second facteur ?
18) Proposer alors une explication quant à la position des groupes d'observations préalablement identifiés, en fonction des variables.

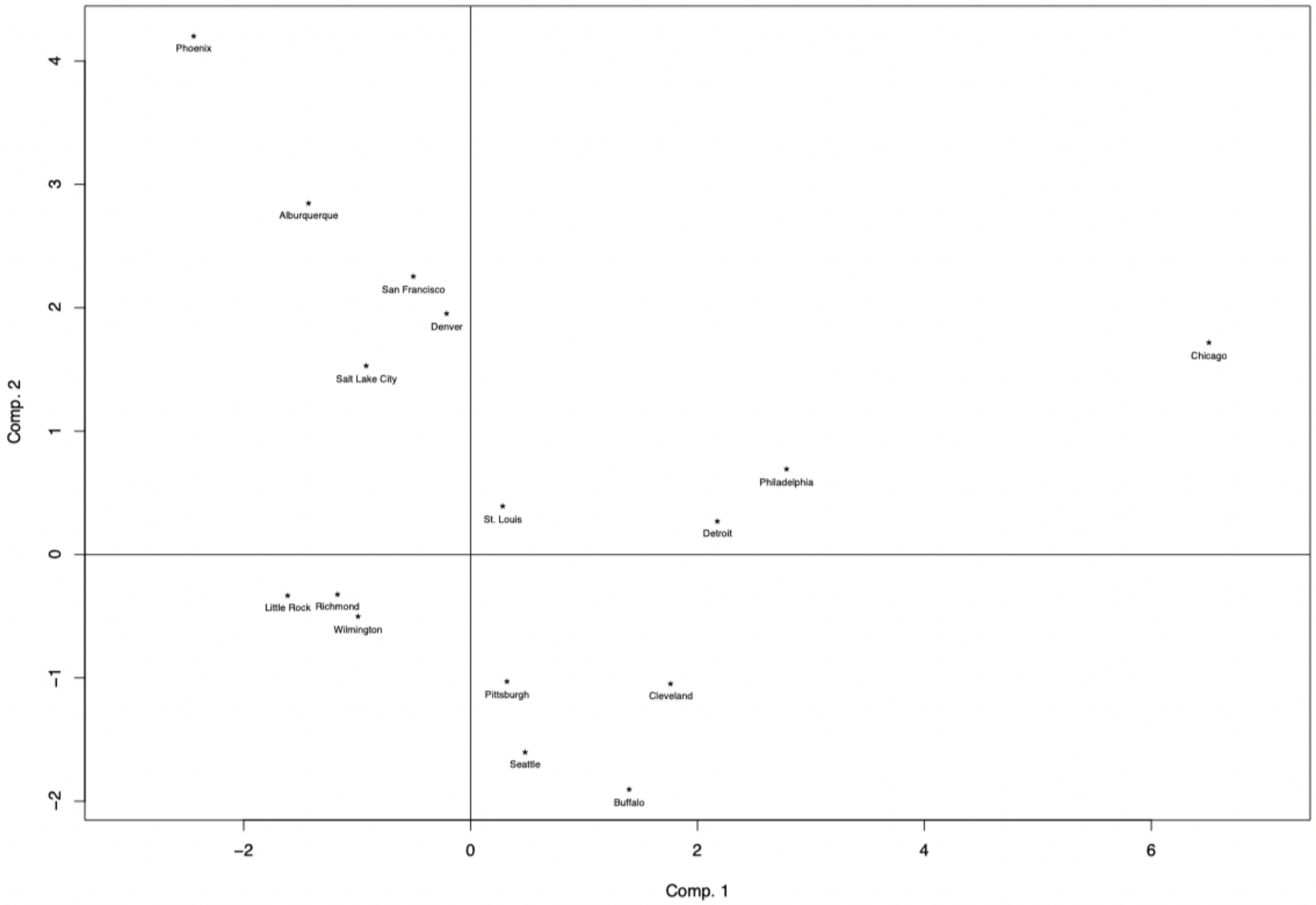

Fig. 4 : observations dans le plan factoriel

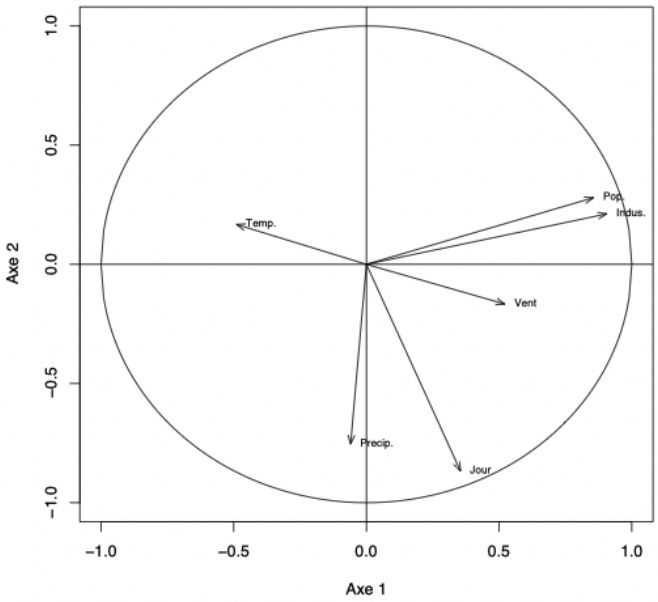

Fig. 5 : variables dans le plan factoriel

Annexe 1

	SO2	Temp.	Indus.	Pop.	Vent	Precip.	Jour
Phoenix	10	21.3	213	582	9.7	17.91	36
Little Rock	13	16.1	91	132	13.2	123.24	100
San Francisco	12	13.7	453	716	14.0	52.48	67
Denver	17	11.1	454	515	14.5	32.89	86
Hartford	56	9.5	412	158	14.5	110.16	127
Wilmington	36	12.2	80	80	14.5	102.23	114
Washington	29	14.1	434	757	15.0	98.78	111
Jacksonville	14	20.2	136	529	14.2	138.35	116
Miami	10	24.2	207	335	14.5	151.89	128
Atlanta	24	16.4	368	497	14.6	122.78	115
Chicago	110	10.3	3344	3369	16.7	87.48	122
Indianapolis	28	11.3	361	746	15.6	98.40	121
Des Moines	17	9.4	104	201	18.0	78.36	103
Wichita	8	13.7	125	277	20.4	77.67	82
Louisville	30	13.1	291	593	13.4	109.50	123
New Orleans	9	20.2	204	361	13.5	144.20	113
Baltimore	47	12.8	625	905	15.4	104.93	111
Detroit	35	9.9	1064	1513	16.3	78.64	129
Minneapolis-St. Paul	29	6.4	699	744	17.1	65.89	137
Kansas City	14	12.5	381	507	16.1	93.98	99
St. Louis	56	13.3	775	622	15.3	91.16	105
Omaha	14	10.8	181	347	17.5	76.66	98
Alburquerque	11	13.8	46	244	14.3	19.74	58
Albany	46	8.7	44	116	14.2	84.73	135
Buffalo	11	8.4	391	463	20.0	91.72	166
Cincinnati	23	12.2	462	453	11.4	99.16	132
Cleveland	65	9.8	1007	751	17.5	88.87	155
Columbus	26	10.8	266	540	13.8	94.01	134
Philadelphia	69	12.6	1692	1950	15.4	101.42	115
Pittsburgh	61	10.2	347	520	15.1	92.00	147
Providence	94	10.0	343	179	17.1	108.59	125
Memphis	10	16.4	337	624	14.8	124.71	105
Nashville	18	15.2	275	448	12.7	116.84	119
Dallas	9	19.0	641	844	17.5	91.29	78
Houston	10	20.5	721	1233	17.4	122.40	103
Salt Lake City	28	10.6	137	176	14.0	38.53	89
Norfolk	31	15.2	96	308	17.1	113.49	116
Richmond	26	14.3	197	299	12.2	108.18	115
Seattle	29	10.6	379	531	15.1	98.53	164
Charleston	31	12.9	35	71	10.5	103.50	148
Milwaukee	16	7.6	569	717	19.0	73.84	123

Annexe 2

	Psi1	**Psi2**	**Cr1**	**Cr2**	**Qual1**	**Qual2**	**Qual12**
Phoenix	-2.4389	4.2016	6.6105	28.6901	23.2107	68.8853	92.0960
Little Rock	-1.6096	-0.3327	2.8794	0.1798	69.0330	2.9485	71.9816
San Francisco	-0.5028	2.2505	0.2810	8.2308	4.6314	92.7779	97.4093
Denver	-0.2101	1.9456	0.0490	6.1521	0.7694	66.0062	66.7756
Hartford	-0.2143	-0.9853	0.0510	1.5776	2.1212	44.8283	46.9495
Wilmington	-0.9908	-0.5097	1.0909	0.4222	62.2975	16.4884	78.7859
Washington	-0.0214	0.0591	0.0005	0.0057	0.2639	2.0065	2.2704
Jacksonville	-1.2199	-0.8269	1.6539	1.1111	24.8585	11.4199	36.2784
Miami	-1.5334	-1.3694	2.6130	3.0476	18.9347	15.1018	34.0365
Atlanta	-0.6065	-0.5706	0.4088	0.5292	20.7022	18.3247	39.0269
Chicago	6.5103	1.7121	47.1029	4.7640	82.7968	5.7264	88.5232
Indianapolis	0.3056	-0.3629	0.1038	0.2141	20.5467	28.9794	49.5261
Des Moines	-0.1333	0.0347	0.0197	0.0020	0.4672	0.0317	0.4989
Wichita	-0.2099	0.6578	0.0490	0.7033	0.5683	5.5809	6.1491
Louisville	-0.4143	-0.5372	0.1908	0.4691	15.0104	25.2352	40.2455
New Orleans	-1.4590	-0.8716	2.3657	1.2346	30.2796	10.8058	41.0854
Baltimore	0.4999	-0.0194	0.2777	0.0006	46.6043	0.0703	46.6746
Detroit	2.1781	0.2673	5.2722	0.1162	91.6626	1.3810	93.0437
Minneapolis-St. Paul	1.5068	-0.2722	2.5231	0.1204	41.1340	1.3426	42.4766
Kansas City	-0.1311	0.2463	0.0191	0.0986	3.0304	10.6927	13.7232
St. Louis	0.2857	0.3869	0.0907	0.2433	18.6083	34.1297	52.7380
Omaha	-0.1389	0.3658	0.0214	0.2175	0.7554	5.2419	5.9973
Alburquerque	-1.4260	2.8470	2.2597	13.1730	17.0489	67.9631	85.0120
Albany	-0.5332	-0.8121	0.3159	1.0717	8.0683	18.7177	26.7861
Buffalo	1.3981	-1.9085	2.1722	5.9192	19.5594	36.4467	56.0061
Cincinnati	-0.5071	-0.4786	0.2858	0.3723	7.4352	6.6222	14.0574
Cleveland	1.7642	-1.0466	3.4589	1.7802	59.0608	20.7861	79.8469
Columbus	-0.1199	-0.6439	0.0160	0.6738	0.9747	28.0995	29.0742
Philadelphia	2.7856	0.6877	8.6236	0.7687	74.0343	4.5127	78.5470
Pittsburgh	0.3223	-1.0351	0.1154	1.7412	4.6350	47.8022	52.4372
Providence	0.0771	-1.0504	0.0066	1.7933	0.2467	45.7964	46.0431
Memphis	-0.5753	-0.3114	0.3678	0.1576	16.7886	4.9187	21.7073
Nashville	-0.9086	-0.5306	0.9175	0.4576	35.4685	12.0966	47.5650
Dallas	-0.0175	1.2255	0.0003	2.4408	0.0057	28.2333	28.2391
Houston	0.5088	0.1385	0.2877	0.0312	3.7708	0.2793	4.0501
Salt Lake City	-0.9170	1.5258	0.9345	3.7834	14.0627	38.9318	52.9945
Norfolk	-0.5862	-0.7555	0.3818	0.9275	16.0727	26.6976	42.7702
Richmond	-1.1722	-0.3266	1.5270	0.1733	52.8840	4.1045	56.9885
Seattle	0.4817	-1.6029	0.2579	4.1756	5.5718	61.6993	67.2711
Charleston	-1.4199	-1.2090	2.2405	2.3754	26.5740	19.2656	45.8396
Milwaukee	1.3930	-0.1834	2.1564	0.0547	35.7087	0.6189	36.3275